From Black-box to Causal-box: Towards Building More Interpretable Models

Inwoo Hwang **Yushu Pan** **Elias Bareinboim**
Causal Artificial Intelligence Lab
Columbia University
inwoo.hwang@columbia.edu {yushupan, eb}@cs.columbia.edu

Abstract

Understanding the predictions made by deep learning models remains a central challenge, especially in high-stakes applications. A promising approach is to equip models with the ability to answer counterfactual questions – hypothetical "what if?" scenarios that go beyond the observed data and provide insight into a model reasoning. In this work, we introduce the notion of causal interpretability, which formalizes when counterfactual queries can be evaluated from a specific class of models and observational data. We analyze two common model classes – blackbox and concept-based predictors – and show that neither is causally interpretable in general. To address this gap, we develop a framework for building models that are causally interpretable by design. Specifically, we derive a complete graphical criterion that determines whether a given model architecture supports a given counterfactual query. This leads to a fundamental tradeoff between causal interpretability and predictive accuracy, which we characterize by identifying the unique maximal set of features that yields an interpretable model with maximal predictive expressiveness. Experiments corroborate the theoretical findings.

1 Introduction

Despite the remarkable success of deep learning models across a wide range of tasks – including image recognition [7, 16], natural language processing [3, 35], and reinforcement learning [32, 34] – these models remain fundamentally opaque. Although they are highly effective at predicting labels based on statistical correlations in the data, they lack the capacity to explain the reasoning behind their predictions, earning them the colloquial label of "black boxes." In other words, current models are difficult to interpret: they lack the ability to justify why a particular decision was made, identify which input factors were most influential, or reason about how outcomes might differ under alternative, counterfactual conditions. This interpretability gap raises concerns in high-stakes domains such as healthcare, law, and scientific discovery, where understanding how and why a model makes a decision is as important as the decision itself.

A rich body of research on explainable AI (XAI) has been developed to better understand the behavior of learned models. For instance, post-hoc explanation methods such as LIME [30], SHAP [20], and Grad-CAM [31] generate local or visual attributions in terms of pixels or extracted features to help interpret predictions. Other approaches aim to build intrinsically interpretable models, such as those that impose sparsity constraints [22], restrict final layers [38], or leverage decision tree structures [37], often trading off model complexity for greater transparency. While these techniques offer useful insights, they fail to bridge the gap between low-level features and high-level, human-understandable features that might explain the behavior of a model.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Mechanistic Interpretability.

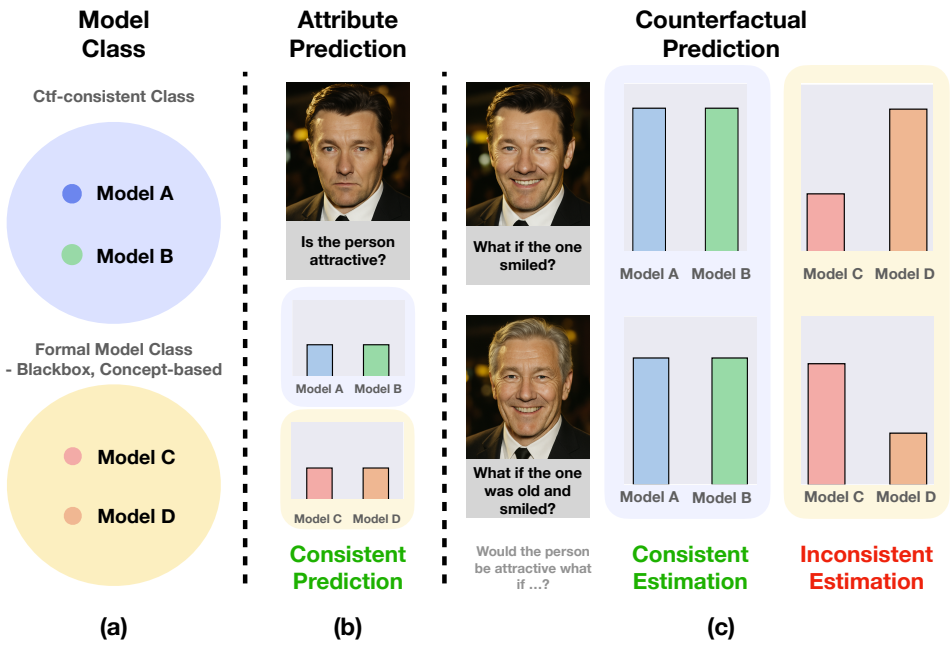


Figure 1: (a) Illustration of different model classes: counterfactually consistent models (blue) and blackbox/concept-based models (yellow). (b) Original input image and corresponding predictions from each model. (c) Counterfactual predictions: models in the top row predict consistently across instantiations within the class, while those in the bottom row produce inconsistent predictions.

One promising avenue for bridging this gap is counterfactual reasoning. Answering what if questions – such as "Would the diagnosis have changed if a different treatment had been administered?" or "Would the person have been classified differently if their income were higher?" – plays a central role in human reasoning and forms the basis of many explanatory and decision-making processes [1, 26, 27]. Enabling AI systems to reason counterfactually opens the door to more interpretable models – ones that can not only predict outcomes accurately but also explain their decisions in a meaningful, human-aligned way.

Recently, concept-based prediction models [15, 23] have been proposed to improve interpretability by enabling reasoning over human-understandable features. These models aim to answer counterfactual queries of the form: "Given an input $\mathbf{x}$, how would the model's prediction change if a feature $\mathbf{W}$ were modified from $\mathbf{w}$ to $\mathbf{w}'$?" Such queries allow users to explore the influence of high-level features – like the presence of a smile or the existence of a tumor – on a model's prediction, providing a possible route to assess whether the model reasoning aligns with human expectations.

Despite their appeal, existing concept-based approaches are oblivious to the causal relationships between features. As a result, they may not reflect the real-world mechanisms or incorporate common-sense knowledge faithfully. While some recent methods attempt to introduce causal structure into concept-based models [4], they frequently lack guarantees of counterfactual consistency – that is, the property that models within the exact class yield consistent answers to the same counterfactual query.

To illustrate this limitation, consider a task of predicting facial attractiveness. Suppose two models, C and D, from the same concept-based class, represented by the yellow circle in Fig. 1-(a), are trained on the same dataset. They first will have the identical attribute prediction, for example, both will predict a lower attractiveness score for the given image (Fig. 1-(c), yellow). However, when they evaluate the counterfactual question "What would the attractiveness be had the person smiled?", model C will maintain the low attractiveness score while model D will raise the attractiveness score (Fig. 1-(c), yellow). This discrepancy reveals a deeper issue: the model class is not counterfactually interpretable, as it does not constrain the space of counterfactual responses. In such cases, users have no principled way to determine which answer to trust, rendering the query effectively unanswerable. In contrast, the model class in blue is desirable since any pair of models – such as model A and B – will give the exact same answer for both attribute and counterfactual predictions. In this case, one

can assert that the attractiveness would be raised had the person smiled, which indicates the model made the decision based on the feature "Smile" and this is aligned with human understanding [8].

In this work, we introduce the notion of causal interpretability, which concerns whether a prediction model can be interpreted consistently across counterfactual scenarios – drawing a connection between XAI and causal inference [1, 26]. Intuitively, a model class is said to be *causally interpretable* if all models within the class yield consistent predictions under counterfactual interventions, as illustrated in blue in Fig. 1. We then show that a blackbox model, which maps inputs directly to labels, is never causally interpretable. That is, such models fundamentally lack the structure needed to answer counterfactual questions. We also demonstrate theoretically that concept-based models [15], which rely on all observed features for prediction, are also not guaranteed to be causally interpretable. Interestingly, we show that causal interpretability can be recovered by constraining the model to use only a certain subset of features.

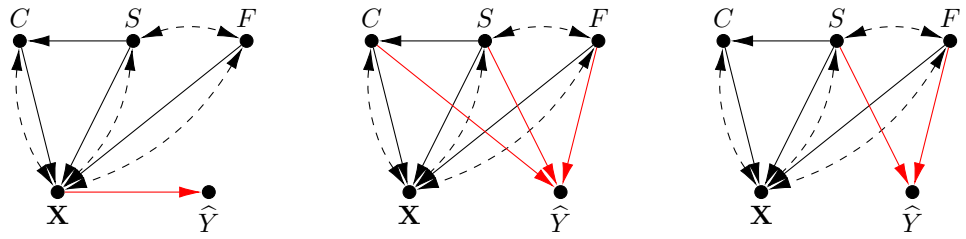
Against this background, we develop a general approach for building causally interpretable models that can answer counterfactual queries consistently by design. Specifically, we propose a complete graphical criterion for determining whether a model that uses a given set of features for prediction is causally interpretable with respect to a counterfactual query. This enables the understanding of (i) which counterfactual questions a given model can answer, and (ii) which models can answer a given counterfactual question. Our framework also reveals a fundamental tradeoff between causal interpretability and predictive accuracy. We characterize the unique maximal set of features that preserves causal interpretability, thereby providing a principled method for building models with maximal expressive power under interpretability constraints. A notable practical implication is that our approach does not require full specification of the causal graph or modeling of unobserved confounders; it only involves the descendants of the target features in the counterfactual query. Experimental results corroborate the proposed theory. More specifically, our contributions are as follows:

- (Sec. 2) We introduce the notion of causal interpretability (Def. 2), which states whether we can evaluate the prediction of the model under counterfactual conditions from observational data. Based on this formulation, we show that a blackbox model is never interpretable (Prop. 1), whereas a concept-based model is also not interpretable in general, in contrast to prior belief.
- (Sec. 3) We develop a graphical criterion that determines whether the model is causally interpretable with respect to the query (Thm. 1). We characterize the unique maximal set of features yielding interpretable architecture (Thm. 2) and provide a practical way of evaluating such queries from the data (Thm. 3). Finally, these results reveal a fundamental tradeoff between the causal interpretability and predictive accuracy (Thm. 4).

Preliminary. Here, we introduce notations and terminologies used in the paper. We use bold letters to denote a set of random variables or their assignments. We use capital letters to denote a random variable or a random vector (e.g., $\mathbf{X}$) and lower case letters to denote their assignments (e.g., $\mathbf{x}$). $\mathbf{x} \cup \mathbf{Z}$ denotes the subset of $\mathbf{x}$ corresponding to variables in $\mathbf{Z}$ and $\mathbf{x} \setminus \mathbf{Z}$ denotes the value of $\mathbf{X} \setminus \mathbf{Z}$ consistent with $\mathbf{x}$.

We employ a structural causal model [1, 26] as our semantical framework. A structural causal model (SCM) $\mathcal{M}$ is a 4-tuple $\langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) \rangle$, where $\mathbf{U}$ is a set of exogenous variables, $\mathbf{V} = \{V_1, \cdots, V_n\}$ is a set of endogenous variables, $\mathcal{F} = \{f_{V_1}, \cdots f_{V_n}\}$ is a set of functions determining $\mathbf{V}$ as $V_j \leftarrow f_{V_j}(\mathbf{Pa}_{V_j}, \mathbf{U}_{V_j})$, where $\mathbf{Pa}_{V_j} \subseteq \mathbf{V} \setminus \{V_j\}$ and $\mathbf{U}_{V_j} \subseteq \mathbf{U}$ for all $V_j \in \mathbf{V}$, and $P(\mathbf{U})$ is a distribution over $\mathbf{U}$. An SCM $\mathcal{M}$ induces a causal diagram $\mathcal{G}$ and a distribution over the endogenous $P(\mathbf{V})$. We use graphical kinship to represent the relationships between the variables. $ND(W)$ denotes non-descendants of a variable W, and $ND(\mathbf{W}) := \cap_{W_i \in \mathbf{W}} ND(W_i)$ denotes non-descendants of a set of variables $\mathbf{W}$. We now define an SCM that describes a generative process that includes images $\mathbf{X}$ and labels prediction $\widehat{Y}$ [24].

Definition 1 (Augmented SCM)**.** *An augmented SCM (ASCM) over a generative level SCM $\mathcal{M}_0 = \langle \mathbf{U}_0, \mathbf{V}_0, \mathcal{F}_0, P^0(\mathbf{U}_0) \rangle$ is a tuple $\mathcal{M} = \langle \mathbf{U}, \{\mathbf{V}, \mathbf{X}, \widehat{Y}\}, \mathcal{F}, P(\mathbf{U}) \rangle$ such that*
(1) exogenous variables $\mathbf{U} = \{\mathbf{U}_0, \mathbf{U}_{\mathbf{X}}\}$;
(2) $\mathbf{V} = \mathbf{V}_0$ are labeled observed endogenous variables, $\mathbf{X}$ is an m-dimensional mixture variable, and $\widehat{Y}$ is a (predicted) label;
(3) $\mathcal{F} = \{\mathcal{F}_0, f_{\mathbf{X}}, f_{\widehat{Y}}\}$, where $f_{\mathbf{X}}$ maps from (the respective domains of) $\mathbf{V} \cup \mathbf{U}_{\mathbf{X}}$ to $\mathbf{X}$ and a



(a) Blackbox prediction (BP). (b) Concept-based prediction (CP). (c) Generalized CP (GCP).

Figure 2: Causal diagrams for different types of predictive models.

classifier $f_{\widehat{Y}}$ maps from (the respective domains of) the subset of $\{\mathbf{V}, \mathbf{X}\}$ to $\widehat{Y}$; and (4) $P(\mathbf{U}_0) = P^0(\mathbf{U}_0)$.

An ASCM $\mathcal{M}$ represents a sequential generative procedure of latent generative factors (i.e., concepts) $\mathbf{V}$, the image $\mathbf{X}$, and the label prediction $\widehat{Y}$. First, the latent features $\mathbf{V}$ are generated by the underlying $\mathcal{M}_0$. The induced causal diagram $\mathcal{G}_{\mathbf{V}}$ is called a latent causal graph (LCG). The high-dimensional mixture $\mathbf{X}$ (e.g., image) is then generated from $\mathbf{V}$ (and $\mathbf{U}_{\mathbf{X}}$), and subsequently, $\widehat{Y}$ is generated from the subset of $\{\mathbf{V}, \mathbf{X}\}$, where $f_{\widehat{Y}}$ is a classifier that predicts the label. We let $\Omega := \{\mathcal{M} : \text{ASCM over } \mathcal{M}_0\}$ be the space of ASCMs. Omitted proofs are provided in Appendix A.2.

2 Causal Intepretability – Foundations

In this section, we formalize the notion of causal interpretability and examine whether existing approaches could elicit counterfactual questions consistently in a valid manner.

We start by analyzing two important classes of predictive models: *blackbox* and *concept-based* models. As illustrated in Fig. 2a, blackbox prediction (BP) models make a prediction on the label from the image pixels $\mathbf{X}$ (i.e., $f_{\widehat{Y}} : \mathcal{D}(\mathbf{X}) \to \mathcal{D}(\widehat{Y})$). In contrast, concept-based prediction (CP) models predict the label based on the generative factors of the image (i.e., $f_{\widehat{Y}} : \mathcal{D}(\mathbf{V}) \to \mathcal{D}(\widehat{Y})$), as illustrated in Fig. 2b. In other words, the classifier of a concept-based model uses the features to make the predictions, instead of the image itself. Formally, a class of BP models and a class of CP models are respectively denoted as Ω_{BP} and Ω_{CP}, where $\Omega_{\text{BP}} := \{\mathcal{M} \in \Omega \mid f_{\widehat{Y}} : \mathcal{D}(\mathbf{X}) \to \mathcal{D}(\widehat{Y})\}$ and $\Omega_{\text{CP}} := \{\mathcal{M} \in \Omega \mid f_{\widehat{Y}} : \mathcal{D}(\mathbf{V}) \to \mathcal{D}(\widehat{Y})\}$. The following examples illustrate the generative process of BP and CP models.

Example 1 (Blackbox Model)**.** *Consider a task of estimating the attractiveness of a human face represented in an image $\mathbf{X}$. Augmented generative process (ASCM) of the prediction by a BP model is given as $\mathcal{M}_{BP} = \langle \mathbf{U} = \{U_F, U_S, U_{C_1}, U_{C_2}, \mathbf{U}_{\mathbf{X}}\}, \{\{F, S, C\}, \mathbf{X}, \widehat{Y}\}, \mathcal{F}^{BP}, P^{BP}(\mathbf{U})\rangle$, where*

$$\mathcal{F}^{BP} = \begin{cases} F \leftarrow U_F \oplus U_S \\ S \leftarrow U_S \\ C \leftarrow (\neg S \wedge U_{C_1}) \oplus (S \wedge U_{C_2}) \\ \mathbf{X} \leftarrow f_{\mathbf{X}}(F, S, C, \mathbf{U}_{\mathbf{X}}) \\ \widehat{Y} \leftarrow f_{\widehat{Y}}(\mathbf{X}), \end{cases} \tag{1}$$

$\widehat{Y}$ is the label (attractiveness) prediction, the exogenous variables $U_F, U_S, U_{C_1}, U_{C_2}$ are independent binary variables, and $P^{BP}(U_F = 1) = 0.4$, $P^{BP}(U_S = 1) = 0.6$, $P^{BP}(U_{C_1} = 1) = 0.3$, $P^{BP}(U_{C_2} = 1) = 0.6$. The exogenous variable $\mathbf{U}_{\mathbf{X}}$ (representing other generative factors) can include (or be correlated to) $\{U_F, U_S, U_{C_1}, U_{C_2}\}$. The causal diagram induced by $\mathcal{M}_{BP}$ is shown in Fig. 2a.

In terms of prediction, the process of obtaining $\widehat{Y}$ has three steps. First, latent generative features F (gender), S (smiling), and C (high cheekbones) are generated. Then, $f_{\mathbf{X}}$ maps the observed generative features $\{F, S, C\}$ and unobserved generative factors $\mathbf{U}_{\mathbf{X}}$ to the images $\mathbf{X}$ in the pixel levels. Finally, the predictor $f_{\widehat{Y}}$ takes these pixels as input to estimate $\widehat{Y}$ in the corresponding model. The functions $f_{\mathbf{X}}$ and $f_{\widehat{Y}}$ can be aggregated as $\widehat{Y} \leftarrow f_{\widehat{Y}} \circ f_{\mathbf{X}}(F, S, C, \mathbf{U}_{\mathbf{X}})$. This illustrates that the

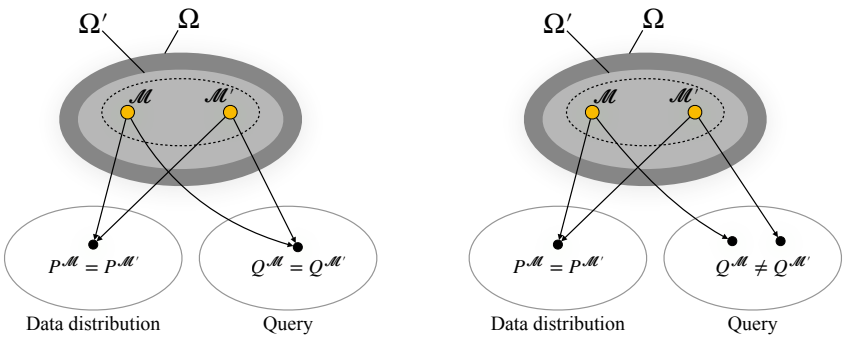


Figure 3: (**Left**) Ω' is causally interpretable if a query can be uniquely computed from the observational data. (**Right**) A query cannot be uniquely computed from the observational data if Ω' is not causally interpretable.

prediction of $\widehat{Y}$ by a BP model is made based on all observed features $\{F, S, C\}$ and unobserved features $\mathbf{U_X}$. ■

Example 2 (Concept-based Model)**.** *The main difference between the class of CP models Ω_{CP} and the class of BP models Ω_{BP} is the form of the classifier $f_{\widehat{Y}}$. Consider the same generative process of observed features $\mathbf{V}_0 = \{F, S, C\}$*[1] *and the image $\mathbf{X}$ in Ex. 1. Let us consider a CP model $\mathcal{M}_{CP} = \langle \mathbf{U} = \{U_F, U_S, U_{C_1}, U_{C_2}, \mathbf{U_X}\}, \{\{F, S, C\}, \mathbf{X}, \widehat{Y}\}, \mathcal{F}^{CP}, P^{CP}(\mathbf{U})\rangle$, where the generative process of $F, S, C, \mathbf{X}$ is the same as Eq. (1), $\widehat{Y}$ is generated as*

$$\widehat{Y} \leftarrow f_{\widehat{Y}}(F, S, C), \tag{2}$$

and $P^{CP}(\mathbf{U})$ is equal to $P^{BP}(\mathbf{U})$ in Ex. 1. In words, this means that instead of predicting $\widehat{Y}$ based on pixels (i.e., image $\mathbf{X}$), the classifier $f_{\widehat{Y}}$ directly predicts $\widehat{Y}$ based on observed features F, S, C. The causal diagram induced by $\mathcal{M}_{CP}$ is shown in Fig. 2b. ■

Examples 1 and 2 illustrate two different types of predictive models, where the classifier predicts the label directly from the image $\mathbf{X}$ (i.e., Ω_{BP}) or from the generative features $\mathbf{V}$ (i.e., Ω_{CP}). While both types have showcased their capability to achieve reasonably high predictive accuracy in many domains [5, 7, 10, 11, 14–16, 23, 33, 40], it is unclear at this moment whether we can interpret how they would predict under counterfactual scenarios, such as "*how attractive the person would be had the one been smiling?*". The following notion of *causal interpretability* formally states whether the counterfactual questions can be answered from the model.

Definition 2 (Causal Interpretability)**.** *Consider a specific model class $\Omega' \subset \Omega$, where Ω is the space of ASCMs. We say the class Ω' is* ***causally interpretable w.r.t. a query*** *Q if $Q^{\mathcal{M}_1} = Q^{\mathcal{M}_2}$ for $\forall \mathcal{M}_1, \mathcal{M}_2 \in \Omega'$ s.t. $P^{\mathcal{M}_1}(\mathbf{V}, \mathbf{X}, \widehat{Y}) = P^{\mathcal{M}_2}(\mathbf{V}, \mathbf{X}, \widehat{Y})$.*

In words, Ω' denotes a certain design choice of the models for predicting the label, that is, it is a space of prediction model candidates (i.e., model class). Ω', for instance, can be Ω_{BP}, when we want to predict the label directly from the image (Fig. 2a), or Ω_{CP}, when the classifier uses all observed features (Fig. 2b). For a query Q, we are concerned with the counterfactual questions such as "*What if the person had smiled?*", which is written in counterfactual notion as $P(\widehat{Y}_{S=1} \mid \mathbf{X} = \mathbf{x})$, and more generally as $Q(\mathbf{W}) := P(\widehat{Y}_{\mathbf{W}} \mid \mathbf{X})$.[2]

In other words, the notion of causal interpretability states whether one can understand the behavior of the model under different counterfactual conditions. If the model is causally interpretable, the counterfactuals can be evaluated from the observational data (Fig. 3, left). Otherwise, the model

[1] In practice, the annotations of the features are provided in many real-world datasets across various domains, e.g., human face [19], medical images [21], and animal species [36]. Otherwise, the common practice is to extract their annotations with vision-language models [29], which is shown to be effective [23, 39].

[2] Note that the definition is general in terms of the query Q, which could vary across different domains, e.g., natural direct effect in fairness analysis [28].

fundamentally cannot answer the counterfactual question from observational data, and thus, we cannot interpret their behavior under counterfactual scenarios (Fig. 3, right). We now analyze two types of predictive models discussed above (i.e., BP model in Ex. 1 and CP model in Ex. 2) and examine their causal interpretability, i.e., whether they can evaluate counterfactuals from observational data.

Example 3 (Continued from Ex. 1)**.** *Consider the BP model $\mathcal{M}_{BP}$ in Ex. 1. Let $\mathbf{U_X}$ includes another independent variable U_S, namely, $\mathbf{U_X} = \{U_S, \mathbf{U_x^-}\}$, where $U_S \perp \mathbf{U} \setminus U_S$; let the observational quantity $P(F = 0, S = 1, C = 1 \mid \mathbf{X} = \mathbf{x}) = 1$, which means that the face is of a male ($F = 0$), who is smiling ($S = 1$), and with the cheekbones high ($C = 1$), given in an image $\mathbf{X} = \mathbf{x}$. The generative process of $\widehat{Y}$ is as $\widehat{Y} \leftarrow f_{\widehat{Y}} \circ f_{\mathbf{X}}(F, S, C, \mathbf{U_X}) = \mathbf{1}[S > 0.5]$.*

Consider another BP model $\mathcal{M}'_{BP}$ with the same generative process of $\mathcal{M}_{BP}$, but for in $\mathcal{M}'_{BP}$, the classifier $f'_{\widehat{Y}}$ is given by: $\widehat{Y} \leftarrow f_{\widehat{Y}} \circ f_{\mathbf{X}}(F, S, C, \mathbf{U_X}) = \mathbf{1}[U_S > 0.5]$. Since $S = U_S$, the two BP models $\mathcal{M}_{BP}$ and $\mathcal{M}'_{BP}$ agrees with the observational data, i.e., $P^{\mathcal{M}_{BP}}(\mathbf{V}, \mathbf{X}, \widehat{Y}) = P^{\mathcal{M}'_{BP}}(\mathbf{V}, \mathbf{X}, \widehat{Y})$, which will lead to the same predictions (and corresponding accuracy).

Now, consider the counterfactual quantity "Given the image $\mathbf{X} = \mathbf{x}$, would the prediction still be attractive ($\widehat{Y} = 1$) had the person not smiled ($S = 0$)?", namely, $Q(S) = P(\widehat{Y}_{S=0} = 1 \mid \mathbf{X} = \mathbf{x})$. Intuitively, a smaller value of $P(\widehat{Y}_{S=0} = 1 \mid \mathbf{X} = \mathbf{x})$ implies the model is more reliable since changing a face to non-smiling reduces the attractiveness in general based on common sense knowledge [8]. For the first BP model $\mathcal{M}_{BP}$, $Q(S)$ evaluates as $P^{\mathcal{M}_{BP}}(\widehat{Y}_{S=0} = 1 \mid \mathbf{X} = \mathbf{x}) = \mathbf{1}[S = 0 > 0.5] = 0$. However, for the second BP model $\mathcal{M}'_{BP}$, $Q(S)$ evaluates as $P^{\mathcal{M}'_{BP}}(\widehat{Y}_{S=0} = 1 \mid \mathbf{X} = \mathbf{x}) = \mathbf{1}[U_S = 1 > 0.5] = 1$. Details for these derivations are provided in Appendix A.

Note that each BP model evaluates the counterfactual query in a completely different way, and the two models are somewhat inconsistent. In practice, if one chooses the class of BP models Ω_{BP} for this prediction task, the above counterfactual question cannot be answered correctly, since two BP models can give an exact opposite answer even if the two models agree perfectly with the observational distribution and their predictions. In other words, the blackbox model class cannot answer counterfactual $Q(S)$ consistently from observational data, and its behavior cannot be interpreted under corresponding counterfactual conditions. ■

One may surmise that Ex. 3 is a pathological case, which for some reason does not allow the evaluation of counterfactual queries in a consistent manner. The next result shows that this is not the case for an arbitrary query $Q(\mathbf{W})$ and a latent causal graph $\mathcal{G}_{\mathbf{V}}$.

Proposition 1 (Non-interpretability of BP)**.** *For any latent causal graph $\mathcal{G}_{\mathbf{V}}$, Ω_{BP} is not causally interpretable w.r.t. $Q(\mathbf{W})$ for any $\mathbf{W} \subseteq \mathbf{V}$.*

Given this impossibility results for the class of blackbox models, one may be tempted to believe that a CP architecture is causally interpretable, as it predicts the label directly from the features where the unobserved factors $\mathbf{U_X}$ are filtered out. However, the following illustrates that this is not the case.

Example 4 (Continued from Ex. 2)**.** *Consider the CP model $\mathcal{M}_{CP}$ in Ex. 2. Similar to Ex. 3, consider an observational quantity $P(F = 0, S = 1, C = 1 \mid \mathbf{X} = \mathbf{x}) = 1$. $\widehat{Y}$ is generated as follows:*

$$\widehat{Y} \leftarrow f_{\widehat{Y}}(F, S, C) = \mathbf{1}[S + C > 0.5]. \tag{3}$$

Now consider another CP model $\mathcal{M}'_{CP}$ that is the same as $\mathcal{M}_{CP}$, except for $C \leftarrow f'_C(S, U_{C_1}) = (S \vee U_{C_1}) \wedge U_{C_2}$ and $P(U_{C_1} = 1) = 0.5$. We have $P^{\mathcal{M}_{CP}}(\mathbf{V}, \mathbf{X}, \widehat{Y}) = P^{\mathcal{M}'_{CP}}(\mathbf{V}, \mathbf{X}, \widehat{Y})$ and $\mathcal{M}'_{CP}$ is compatible with the graphical constraints in Fig. 2b. Now consider the same counterfactual quantity $P(\widehat{Y}_{S=0} = 1 \mid \mathbf{X} = \mathbf{x})$ in Ex. 3. For $\mathcal{M}_{CP}$, we have $P^{\mathcal{M}_{CP}}(\widehat{Y}_{S=0} = 1 \mid \mathbf{X} = \mathbf{x}) = P^{\mathcal{M}_{CP}}(C_{S=0} = 1 \mid F = 0, S = 1, C = 1) = 0.3$. However, for the second CP model, $P^{\mathcal{M}'_{CP}}(\widehat{Y}_{S=0} = 1 \mid \mathbf{X} = \mathbf{x}) = P^{\mathcal{M}'_{CP}}(C_{S=0} = 1 \mid F = 0, S = 1, C = 1) = 0.5$. This implies that the two CP models are also inconsistent w.r.t $Q(S)$. In other words, even prediction using features $\mathbf{V}$, not pixels $\mathbf{X}$, counterfactual queries induced by the CP models can still differ from each other. ■

3 A Causal Approach Towards More Interpretable Models

In this section, we establish a principled way of understanding causal interpretability from a graphical point of view and propose a generalized framework for building causally interpretable models.

3.1 Generalized Concept-based Models

We first define generalized concept-based prediction (GCP) models, a broader class that predicts the label from an arbitrary set of observed features.

Definition 3 (Generalized Concept-based Prediction)**.** *Let* $\mathbf{T} \subseteq \mathbf{V}$ *be a set of features that is used as a predictor of the label. That is, a classifier* $f_{\widehat{Y}}$ *makes a prediction on a label based on* $\mathbf{T}$*. We say such predictive models as generalized concept-based models. A class of GCP models that employ the features* $\mathbf{T}$ *for prediction is denoted as* $\Omega_{\text{GCP}(\mathbf{T})} := \{\mathcal{M} \in \Omega \mid f_{\widehat{Y}} : \mathcal{D}(\mathbf{T}) \to \mathcal{D}(\widehat{Y})\}$.

Compared to CP models, GCP models employ a selected set of features $\mathbf{T} \subseteq \mathbf{V}$ as a predictor of the label, which relaxes the requirement of CP where all features are considered.

The selection of the features $\mathbf{T}$ in a GCP model should be specified during the model building stage, and our goal is to understand the implications of different choices of $\mathbf{T}$ and which ones could lead to causally interpretable models (i.e., satisfying Def. 2). To answer this question systematically, we introduce a graphical criterion for determining whether a model satisfies causal interpretability.

Theorem 1 (Graphical Criterion)**.** *Consider GCP models that employ a set of features* $\mathbf{T}$ *as a predictor of the label.* $\Omega_{\text{GCP}(\mathbf{T})}$ *is causally interpretable w.r.t. a query* $Q(\mathbf{W})$ *if and only if* $\mathbf{T} \subseteq \mathbf{W} \cup ND(\mathbf{W})$.

In words, this result says that a query $Q(\mathbf{W})$ can be evaluated if the model uses the features among $\mathbf{W}$ or non-descendants of $\mathbf{W}$ to make a prediction on the label. In other words, the models that use any descendant of $\mathbf{W}$ cannot answer counterfactual question and no guarantee can be provided on how they would make predictions under the corresponding counterfactual scenarios.[3]

Thm. 1 enables one to identify the architectures (associated with $\mathbf{T}$) that are causally interpretable with respect to given counterfactual queries. Interestingly, the models that are potentially causally interpretable are not unique. The following formalizes the notion of admissible architectures.

Definition 4 (T-Admissible Set)**.** *We say* $\mathbf{T}$ *is T-admissible w.r.t.* $\mathbf{W}_\star = \{\mathbf{W}_1, \mathbf{W}_2, \cdots\}$ *if* $\Omega_{\text{GCP}(\mathbf{T})}$ *is interpretable w.r.t.* $Q(\mathbf{W}_i)$ *for all* $\mathbf{W}_i \in \mathbf{W}_\star$*. A set of T-admissible sets w.r.t.* $\mathbf{W}_\star$ *is denoted as T-Ad*$(\mathbf{W}_\star)$.

To illustrate, T-admissible set represents model architectures that can answer (potentially multiple) counterfactual queries $Q(\mathbf{W}_1), Q(\mathbf{W}_2), \cdots$. For example, in Fig. 2, eligible models that one can evaluate $Q(S)$ is GCP models whose classifier employs $\{S\}$, $\{F\}$, or $\{S, F\}$ as a predictor of the label, i.e., T-admissible set corresponds to the query $Q(\{S\})$ is T-Ad$(\{S\}) = \{\{S\}, \{F\}, \{S, F\}\}$.

Given the multiplicity of admissible models, our goal is to find the models that use as many features as possible to predict the label $\widehat{Y}$, i.e., maximal $\mathbf{T}$, as it would be beneficial in terms of predictive accuracy. We denote it as a *maximal T-admissible set*, which is formally defined below.

Definition 5 (Maximal T-Admissible Set)**.** *Suppose* $\mathbf{S} \in$ *T-Ad*$(\mathbf{W}_\star)$ *and* $\mathbf{S}' \notin$ *T-Ad*$(\mathbf{W}_\star)$ *for any* $\mathbf{S}' \supsetneq \mathbf{S}$*. We denote such* $\mathbf{S}$ *as Max-T-Ad*$(\mathbf{W}_\star)$.

In other words, a maximal T-admissible set is a T-admissible set that would cease to be T-admissible if any additional variable were added to it. Note that once a set is not T-admissible, adding more variables never makes it T-admissible again by Thm. 1. Identifying a maximal T-admissible set would lead to a model with maximal predictive power while retaining causal interpretability. One might suspect that multiple maximal T-admissible sets could exist, making it unclear which to select to maximize the predictive expressiveness. However, the next result says that this is not the case, since we can establish the uniqueness of the maximal T-admissible set.

Theorem 2 (Uniqueness of Maximal T-Admissible Set)**.** *For the queries* $Q(\mathbf{W}_\star)$*, a maximal T-admissible set is unique and can be written as:*

$$\textit{Max-T-Ad}(\mathbf{W}_\star) = \cap_{\mathbf{W}_i \in \mathbf{W}_\star} \left(\mathbf{W}_i \cup ND(\mathbf{W}_i)\right). \tag{4}$$

Also, $\mathbf{T} \in$ *T-Ad*$(\mathbf{W}_\star)$ *if and only if* $\mathbf{T} \subseteq$ *Max-T-Ad*$(\mathbf{W}_\star)$.

[3]Note that for the case of $\mathbf{X} = \mathbf{T}$, Ω_{BP} is not interpretable w.r.t. any $Q(\mathbf{W})$ since $\mathbf{X}$ is a descendant of $\mathbf{W}$ for any $\mathbf{W} \subseteq \mathbf{V}$, generalizing Prop. 1. Similarly, $\Omega_{\text{GCP}(\mathbf{T})}$ is also never interpretable if $\mathbf{X} \in \mathbf{T}$, i.e., *hybrid* models that make predictions based on the combination of the image and features.

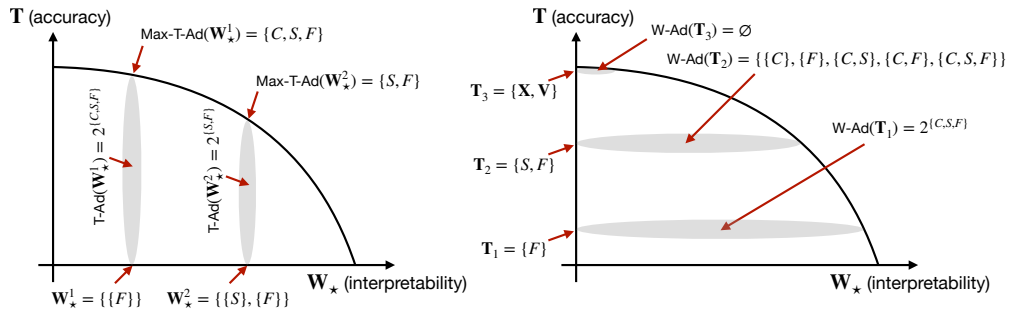


Figure 4: (**Left**) As we want a model to answer more counterfactual queries ($\mathbf{W}^1_\star \subseteq \mathbf{W}^2_\star$), the predictive power would decrease (Max-T-Ad($\mathbf{W}^2_\star$) $\subseteq$ Max-T-Ad($\mathbf{W}^1_\star$)). (**Right**) As the predictive power increases ($\mathbf{T}_1 \subseteq \mathbf{T}_2$), interpretable counterfactuals would decrease (W-Ad($\mathbf{T}_2$) $\subseteq$ W-Ad($\mathbf{T}_1$)).

To illustrate, for the group of queries $Q(\mathbf{W}_1), Q(\mathbf{W}_2), \cdots$, the maximal T-admissible set is unique and it is the intersection of non-descendants of $\mathbf{W}_i$ plus $\mathbf{W}_i$. Interestingly, identifying a maximal T-admissible set only requires the descendants of $\mathbf{W}$ and does not rely on the full specification of the causal graph. For example, given the features {cheekbone, smiling, gender} and the query "What if the person had smiled?", it only requires the knowledge of descendants of "smiling", which is "cheekbone". This does not rely on the full latent causal graph, which is often challenging to obtain.

An important practical implication of Thms 1 and 2 is that, given a query $Q(\mathbf{W})$, one could incorporate additional features as long as they are non-descendants of $\mathbf{W}$, which would help improve accuracy while retaining the causal interpretability w.r.t. $P(\widehat{Y}_{\mathbf{W}} \mid \mathbf{X})$. For example, given the T-admissible set {smiling, gender} and the query "Would the person be attractive had they smiled?", one can incorporate additional features, e.g., age or hair color, that are non-descendants of smiling.

So far, we have described how to find causally interpretable models that can answer counterfactual queries. We now describe a practical way of evaluating such queries from the data.

Theorem 3 (Closed Form)**.** *If $\Omega_{\text{GCP}(\mathbf{T})}$ is causally interpretable w.r.t. $Q(\mathbf{W})$, the following holds:*

$$P(\widehat{Y}_{\mathbf{w}'} \mid \mathbf{x}) = \sum_{\mathbf{t}} P(\widehat{Y} \mid \mathbf{w}' \cap \mathbf{T}, \mathbf{t} \setminus \mathbf{W}) P(\mathbf{t} \mid \mathbf{x}). \tag{5}$$

This implies that the counterfactual quantity can be elicited from a two-step prediction – (1) a classifier $P(\widehat{Y} \mid \mathbf{T})$ and (2) a feature extractor $P(\mathbf{T} \mid \mathbf{X})$. For example, $Q(S)$ introduced in Ex. 3 can be computed using observational data and the maximal T-admissible set {S, F} as: $P(\widehat{Y}_{S=0} \mid \mathbf{X}) = \sum_{s,f} P(\widehat{Y} \mid S = 0, f) P(s, f \mid \mathbf{X})$. Specifically, $\{S, F\}$ are extracted from $P(S, F \mid \mathbf{X})$ and the prediction is made by classifying $P(\widehat{Y} \mid S = 0, F)$, conditioning $S = 0$. Note that Eq. (5) only holds when the model is causally interpretable, and it does not hold for non-interpretable ones.

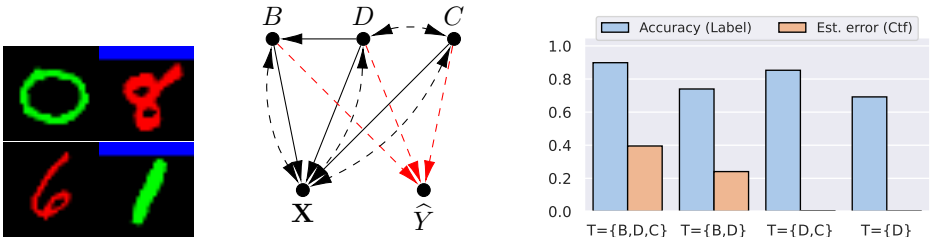
3.2 Fundamental Trade-Off between Causal Interpretability and Accuracy

So far, we have developed the machinery for building causally interpretable models that can answer counterfactual queries. Now, we discuss which queries can be read from the given predictive model architecture. The following formalizes such notions of admissible queries.

Definition 6 (W-Admissible Set)**.** *We say $\mathbf{W}$ is W-admissible w.r.t. $\mathbf{T}$ if $\Omega_{\text{GCP}(\mathbf{T})}$ is causally interpretable w.r.t. $Q(\mathbf{W})$. A set of W-admissible sets w.r.t. $\mathbf{T}$ is denoted as W-Ad($\mathbf{T}$).*

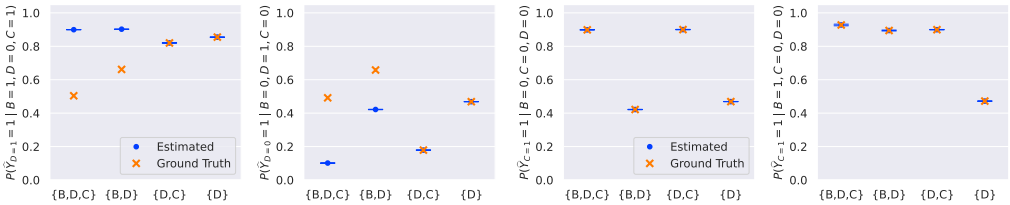
For example, in Fig. 2b, CP model that uses the features $\{F, S, C\}$ as the predictor of the label can answer counterfactual queries $Q(\{F\}), Q(\{C\}), Q(\{F, S\}), Q(\{F, C\})$ and $Q(\{F, S, C\})$, i.e., W-Ad($\{F, S, C\}$) $= \{\{F\}, \{C\}, \{F, S\}, \{F, C\}, \{F, S, C\}\}$ by applying Thm. 1. Similarly, in Fig. 2c, we have W-Ad($\{S, C\}$) $= \{\{F\}, \{S\}, \{C\}, \{F, S\}, \{F, C\}, \{S, C\}, \{F, S, C\}\}$. Here, one might notice that the model using a larger set of features can answer a smaller number of counterfactual questions. Our next result establishes a trade-off between accuracy and causal interpretability.

Theorem 4 (Causal Interpretability-Accuracy Trade-Off)**.** *The following holds:*
(i) If $\mathbf{T}_1 \subseteq \mathbf{T}_2$, then W-Ad($\mathbf{T}_2$) $\subseteq$ W-Ad($\mathbf{T}_1$).
(ii) If $\mathbf{W}^1_\star \subseteq \mathbf{W}^2_\star$, then Max-T-Ad($\mathbf{W}^2_\star$) $\subseteq$ Max-T-Ad($\mathbf{W}^1_\star$).



(a) (b) (c)

Figure 5: (a) Example images of BarMNIST dataset. (b) Causal diagram of GCP models. Red arrows represent the possible usage for predicting the label. (c) Causal interpretability-accuracy trade-off.



(a) Changing digit. (b) Changing color.

Figure 6: Estimation of counterfactual queries. Blue dots and orange marks denote estimation of counterfactual queries and ground truth value, respectively.

In other words, Thm. 4-(i) states that the counterfactuals that can be evaluated from the model decrease (W-Ad($\mathbf{T}_2$) $\subseteq$ W-Ad($\mathbf{T}_1$)) as the predictors increase ($\mathbf{T}_1 \subseteq \mathbf{T}_2$). Similarly, Thm. 4-(ii) states that the predictive power would decrease (Max-T-Ad($\mathbf{W}_\star^2$) $\subseteq$ Max-T-Ad($\mathbf{W}_\star^1$)) as we want the models to answer more counterfactual queries ($\mathbf{W}_\star^1 \subseteq \mathbf{W}_\star^2$). This reveals a fundamental trade-off between causal interpretability and accuracy, where better predictive power would compromise the interpretability, and vice versa, as illustrated in Fig. 4.

4 Experiments

In this section, we evaluate our framework for estimating counterfactuals and compare it with prior approaches. Experimental details and additional experimental results are provided in Appendix B.

4.1 Synthetic datasets

We design the BarMNIST dataset [17, 24] where the digits are colored and a bar appears at the top of the image, as shown in Fig. 5a. Specifically, we consider the features "bar" (B), "digit" (D), and "color" (C), where D, C are correlated and D has a direct causal effect on B, as illustrated in Fig. 5b. The true label is generated from all of the features and unobserved factors.

The dataset allows us to compare the estimation of counterfactuals from each model with the ground-truth. We trained 4 different models, each using $\mathbf{T} = \{B, D, C\}$, $\{B, D\}$, $\{D, C\}$, and $\{D\}$ as the predictor of the label. As shown in Fig. 5c, the model using $\mathbf{T} = \{B, D, C\}$ achieves the best accuracy, followed by $\mathbf{T} = \{B, D\}$ and $\mathbf{T} = \{D, C\}$, and the model using $\mathbf{T} = \{D\}$ shows the lowest accuracy. On the other hand, the best model ($\mathbf{T} = \{B, D, C\}$) in terms of accuracy shows a high estimation error on the counterfactual query of changing the digit. Thm. 1 suggests that any estimation using observed data cannot capture the true counterfactual prediction of this model, since it uses B, which is the descendant of D. For the same reason, $\mathbf{T} = \{B, D\}$ is not causally interpretable, in contrast to $\mathbf{T} = \{D, C\}$ and $\mathbf{T} = \{D\}$. Our theory (Thm. 2) also suggests that there exists a unique maximal set of features that maintains causal interpretability, in this case, $\mathbf{T} = \{D, C\}$.

In Fig. 6, we take a closer look at how these models estimate counterfactuals. As shown in Fig. 6a, $\mathbf{T} = \{D, C\}$ and $\mathbf{T} = \{D\}$ are admissible models for the counterfactual query of changing the digit. On the other hand, for changing color (Fig. 6b), all models are admissible and output a correct estimate of the counterfactual query, since C is not a descendant of any other features.

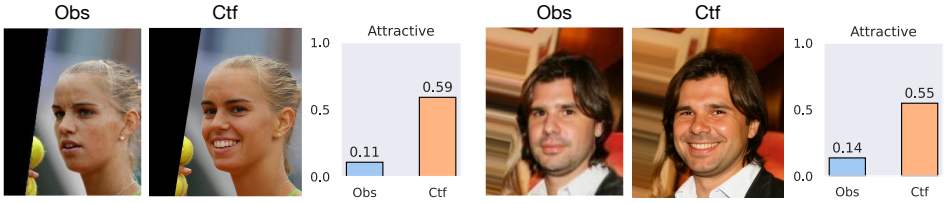


Figure 7: Visualization of interpreting counterfactual predictions on CelebA examples.

4.2 Real-world datasets

CelebA dataset [19] contains human face images with the annotations on facial expressions and attributes, such as "smiling", "age", "gender", etc. We consider a model predicting the label "attractiveness" and examine how a model makes a prediction under counterfactual conditions "Would the person look attractive had they smiled?". In the real world, it is impossible to observe a counterfactual outcome, but our theory allows us to interpret the behavior of (causally interpretable) models under counterfactual conditions. Based on Thm. 1, we choose the features that are not the descendants of smiling. Fig. 7 illustrates the counterfactual prediction of the model using non-descendant features (i.e., "smiling" and "gender"). We can interpret its behavior under the counterfactual condition that it predicts a higher attractiveness had the one smiled, which is aligned with human common sense.

5 Conclusion

In this work, we introduced the notion of causal interpretability, which states whether counterfactual queries can be evaluated from a model and observational data. By examining commonly used model classes – blackbox and concept-based models – we demonstrated that neither is causally interpretable. To this end, we developed a graphical criterion that determines whether the model is causally interpretable with respect to the query (Thm. 1). We characterize the unique maximal set of features yielding interpretable architecture (Thm. 2) and provide a practical way of evaluating such queries from the data (Thm. 3). Our results reveal a fundamental tradeoff between the causal interpretability and predictive accuracy (Thm. 4). Theoretical findings are corroborated by the experimental results. Additional discussions and limitations are provided in Appendix C.

Acknowledgments and Disclosure of Funding

We thank anonymous reviewers for their constructive comments. This research is supported in part by the NSF, ONR, AFOSR, DoE, Amazon, JP Morgan, and The Alfred P. Sloan Foundation.

References

[1] Elias Bareinboim, Juan D. Correa, Duligur Ibeling, and Thomas Icard. On pearl's hierarchy and the foundations of causal inference. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, page 507–556. Association for Computing Machinery, New York, NY, USA, 2022.

[2] Juan Correa, Sanghack Lee, and Elias Bareinboim. Nested counterfactual identification from arbitrary surrogate experiments. *Advances in Neural Information Processing Systems*, 34: 6856–6867, 2021.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[4] Gabriele Dominici, Pietro Barbiero, Mateo Espinosa Zarlenga, Alberto Termine, Martin Gjoreski, Giuseppe Marra, and Marc Langheinrich. Causal concept graph models: Beyond causal opacity in deep learning. *arXiv preprint arXiv:2405.16507*, 2024.

[5] Gabriele Dominici, Pietro Barbiero, Francesco Giannini, Martin Gjoreski, Giuseppe Marra, and Marc Langheinrich. Counterfactual concept bottleneck models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[6] Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Zohreh Shams, Frederic Precioso, Stefano Melacci, Adrian Weller, et al. Concept embedding models: Beyond the accuracy-explainability trade-off. *Advances in Neural Information Processing Systems*, 35:21400–21413, 2022.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[8] Simone Horn, Natalia Matuszewska, Nikolaos Gkantidis, Carlalberta Verna, and Georgios Kanavakis. Smile dimensions affect self-perceived smile attractiveness. *Scientific reports*, 11 (1):2779, 2021.

[9] Inwoo Hwang, Yesong Choe, Yeahoon Kwon, and Sanghack Lee. On positivity condition for causal inference. In *International Conference on Machine Learning*, pages 20818–20841. PMLR, 2024.

[10] Aya Abdelsalam Ismail, Tuomas Oikarinen, Amy Wang, Julius Adebayo, Samuel Don Stanton, Hector Corrada Bravo, Kyunghyun Cho, and Nathan C. Frey. Concept bottleneck language models for protein design. In *The Thirteenth International Conference on Learning Representations*, 2025.

[11] Sujin Jeon, Hyundo Lee, Eungseo Kim, Sanghack Lee, Byoung-Tak Zhang, and Inwoo Hwang. Locality-aware concept bottleneck model. *arXiv preprint arXiv:2508.14562*, 2025.

[12] Yonghan Jung, Jin Tian, and Elias Bareinboim. Estimating identifiable causal effects through double machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12113–12122, 2021.

[13] Yonghan Jung, Iván Díaz, Jin Tian, and Elias Bareinboim. Estimating causal effects identifiable from a combination of observations and experiments. *Advances in Neural Information Processing Systems*, 36:46446–46490, 2023.

[14] Eunji Kim, Dahuin Jung, Sangha Park, Siwon Kim, and Sungroh Yoon. Probabilistic concept bottleneck models. In *International Conference on Machine Learning*, pages 16521–16540. PMLR, 2023.

[15] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pages 5338–5348. PMLR, 2020.

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[17] Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

[18] Adam Li, Yushu Pan, and Elias Bareinboim. Disentangled representation learning in non-markovian causal systems. *Advances in Neural Information Processing Systems*, 37:104843–104903, 2024.

[19] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018):11, 2018.

[20] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[21] Michael Nevitt, David Felson, and Gayle Lester. The osteoarthritis initiative. *Protocol for the cohort study*, 1:2, 2006.

[22] Andrew Ng et al. Sparse autoencoder. *CS294A Lecture notes*, 72(2011):1–19, 2011.

[23] Tuomas Oikarinen, Subhro Das, Lam M. Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. In *The Eleventh International Conference on Learning Representations*, 2023.

[24] Yushu Pan and Elias Bareinboim. Counterfactual image editing. In *International Conference on Machine Learning*, pages 39087–39101. PMLR, 2024.

[25] Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

[26] Judea Pearl. *Causality*. Cambridge university press, 2009.

[27] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.

[28] Drago Plecko and Elias Bareinboim. Causal fairness analysis. *arXiv preprint arXiv:2207.11385*, 2022.

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[30] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[31] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[32] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[33] Chung-En Sun, Tuomas Oikarinen, Berk Ustun, and Tsui-Wei Weng. Concept bottleneck large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[34] Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, volume 30, pages 5998–6008, 2017.

[36] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[37] Alvin Wan, Lisa Dunlap, Daniel Ho, Jihan Yin, Scott Lee, Henry Jin, Suzanne Petryk, Sarah Adel Bargal, and Joseph E Gonzalez. Nbdt: Neural-backed decision trees. *arXiv preprint arXiv:2004.00221*, 2020.

[38] Eric Wong, Shibani Santurkar, and Aleksander Madry. Leveraging sparse linear layers for debuggable deep networks. In *International Conference on Machine Learning*, pages 11205–11216. PMLR, 2021.

[39] Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19187–19197, 2023.

[40] Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. In *The Eleventh International Conference on Learning Representations*, 2023.

Appendix

A Proofs and Additional Examples **13**

A.1 Derivations in Examples 13

A.2 Omitted Proofs 14

A.3 Additional Examples 17

B Experiments **18**

B.1 Dataset 18

B.2 Experimental Details 19

B.3 Additional Experimental Results 19

C Additional Discussions, Limitations, and Future Work **21**

A Proofs and Additional Examples

A.1 Derivations in Examples

A.1.1 Derivation in Ex. 3

In Ex. 3, for the first BP model $\mathcal{M}_{\text{BP}}$, we evaluate $Q(S)$ from $\mathcal{M}_{\text{BP}}$ as follows:

$$
\begin{aligned}
&P^{\mathcal{M}_{\text{BP}}}(\widehat{Y}_{S=0} = 1 \mid \mathbf{X} = \mathbf{x}) \\
=&\sum_{f,s,c} P^{\mathcal{M}_{\text{BP}}}(\widehat{Y}_{S=0} = 1 \mid F = f, S = s, C = c, \mathbf{X} = \mathbf{x}) P^{\mathcal{M}_{\text{BP}}}(F = f, S = s, C = c \mid \mathbf{X} = \mathbf{x}) \\
=&P^{\mathcal{M}_{\text{BP}}}(\widehat{Y}_{S=0} = 1 \mid F = 0, S = 1, C = 1, \mathbf{X} = \mathbf{x}) \\
=&P^{\mathcal{M}_{\text{BP}}}(f_{\widehat{Y}} \circ f_{\mathbf{X}}(F, S, C, \mathbf{U}_{\mathbf{X}})_{S=0} = 0 \mid F = 0, S = 1, C = 1, \mathbf{X} = \mathbf{x}) \\
=&P^{\mathcal{M}_{\text{BP}}}(\mathbf{1}[S > 0.5]_{S=0} = 0 \mid F = 0, S = 1, C = 1, \mathbf{X} = \mathbf{x}), \\
=&\mathbf{1}[S = 0 > 0.5] = 0.
\end{aligned}
$$

However, for the second BP model, we evaluate $Q(S)$ from $\mathcal{M}'_{\text{BP}}$ as:

$$
\begin{aligned}
&P^{\mathcal{M}'_{\text{BP}}}(\widehat{Y}_{S=0} = 1 \mid \mathbf{X} = \mathbf{x}) \\
=&P^{\mathcal{M}'_{\text{BP}}}(f'_{\widehat{Y}} \circ f_{\mathbf{X}}(F, S, C, \mathbf{U}_{\mathbf{X}})_{S=0} = 0 \mid F = 0, S = 1, C = 1, \mathbf{X} = \mathbf{x}) \\
=&P^{\mathcal{M}'_{\text{BP}}}(\mathbf{1}[U_S > 0.5]_{S=0} = 0 \mid F = 0, S = 1, C = 1, \mathbf{X} = \mathbf{x}) \\
=&\sum_{\mathbf{u}} P^{\mathcal{M}'_{\text{BP}}}(\mathbf{1}[U_S > 0.5]_{S=0} = 0 \mid \mathbf{u}) P^{\mathcal{M}'_{\text{BP}}}(\mathbf{u} \mid F = 0, S = 1, C = 1, \mathbf{X} = \mathbf{x})) \quad \text{(summing over } \mathbf{U}) \\
=&P^{\mathcal{M}'_{\text{BP}}}(\mathbf{1}[U_S > 0.5]_{S=0} = 0 \mid U_S = 1) \qquad (S = U_S) \\
=&\mathbf{1}[U_S = 1 > 0.5] = 1.
\end{aligned}
$$

A.1.2 Derivation in Ex. 4

In Ex. 4, for $\mathcal{M}_{\text{CP}}$,

$$
\begin{aligned}
&P^{\mathcal{M}_{\text{CP}}}(\widehat{Y}_{S=0} = 1 \mid \mathbf{X} = \mathbf{x}) \\
=&P^{\mathcal{M}_{\text{CP}}}(\widehat{Y}_{S=0} = 1 \mid F = 0, S = 1, C = 1, \mathbf{X} = \mathbf{x}) \\
=&P^{\mathcal{M}_{\text{CP}}}(\widehat{Y}_{S=0} = 1 \mid F = 0, S = 1, C = 1) && (\widehat{Y} \perp \mathbf{X} \mid \mathbf{V}) \\
=&\sum_{c} P^{\mathcal{M}_{\text{CP}}}(\widehat{Y}_{S=0} = 1 \mid C_{S=0} = c) P^{\mathcal{M}_{\text{CP}}}(C_{S=0} = c \mid F = 0, S = 1, C = 1) \\
=&\sum_{c} P^{\mathcal{M}_{\text{CP}}}(\widehat{Y}_{S=0} = 1 \mid C_{S=0} = c) P^{\mathcal{M}_{\text{CP}}}(C_{S=0} = c \mid F = 0, S = 1, C = 1) \\
=&P^{\mathcal{M}_{\text{CP}}}(C_{S=0} = 1 \mid F = 0, S = 1, C = 1) && (\text{Eq. } 3) \\
=&0.3
\end{aligned}
$$

A.2 Omitted Proofs

In this section, we present the proofs of our theoretical results in Sec. 2 and 3. We first formally introduce the causal diagram induced by an SCM.

Definition 7 (Causal Diagram [1, Def. 13])**.** *Consider an SCM $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) \rangle$. We construct a graph $\mathcal{G}$ using $\mathcal{M}$ as follows:*

(1) add a vertex for every variable in $\mathbf{V}$,

(2) add a directed edge ($V_j \rightarrow V_i$) for every $V_i, V_j \in \mathbf{V}$ if V_j appears as an argument of $f_{V_i} \in \mathcal{F}$,

(3) add a bidirected edge ($V_j \longleftrightarrow V_i$) for every $V_i, V_j \in \mathbf{V}$ if the corresponding $\mathbf{U}_{V_i}, \mathbf{U}_{V_j} \subseteq \mathbf{U}$ are not independent or if f_{V_i} and f_{V_j} share some $U \in \mathbf{U}$ as an argument.

We refer to $\mathcal{G}$ as the causal diagram induced by $\mathcal{M}$ (or "causal diagram of $\mathcal{M}$" for short). ■

We then formally introduce the identifiability of a counterfactual query given an observational distribution and a causal diagram $\mathcal{G}$.

Definition 8 (Counterfactual Identification)**.** *A counterfactual query $P(y_{1[\mathbf{x}_1]}, y_{2[\mathbf{x}_2]}, ...)$ is said to be identifiable from $P(\mathbf{V})$ and $\mathcal{G}$, if $P(y_{1[\mathbf{x}_1]}, y_{2[\mathbf{x}_2]}, ...)$ is uniquely computable from the distributions $P(\mathbf{V})$ in any SCM that induces $\mathcal{G}$.*

Then we start from two lemmas as a tool for the proof of Thm. 1.

Lemma 1. *Consider an SCM $\mathcal{M}$ over $\mathbf{V}$. Suppose that there exists a path made entirely of bi-directed edges between $V_i, V_j \in \mathbf{V}$ in $\mathcal{G}$. Consider two sets $\mathbf{A}, \mathbf{B} \subseteq \mathbf{V}$ and $\mathbf{A} \cap \mathbf{B} = \emptyset$. Let the intervened values are not consistent with the factual values, namely, $\mathbf{b} \not\subseteq \mathbf{v}$. Then the query $P(\mathbf{a_b} \mid \mathbf{v})$ is identifiable from $P(\mathbf{V})$ and $\mathcal{G}$ if and only if $\mathbf{A} \subseteq ND(\mathbf{B})$, where $ND(\mathbf{B}) = \cap_{B_i \in \mathbf{B}} ND(B_i)$.*

Proof. ($\Rightarrow$) Suppose $\mathbf{A} \subseteq ND(\mathbf{B})$. We have $P(\mathbf{a_b} \mid \mathbf{v}) = P(\mathbf{a} \mid \mathbf{v}) = \mathbf{1}[\mathbf{a} = \mathbf{v}]$ which implies that $P(\mathbf{a_b} \mid \mathbf{v})$ is uniquely computable.

($\Leftarrow$) Suppose there exists $A \in \mathbf{A}$ such that $A \in Des(B)$. By Correa et al. [2, Thm. 3], $P(\mathbf{a_b} \mid \mathbf{v})$ is an inconsistent factor since $\mathbf{B} \subseteq \mathbf{V}$ and $\mathbf{b} \subseteq \mathbf{v}$, and thus, it is not identifiable from $P(\mathbf{V})$. □

Lemma 2 (Correa et al. [2, Lemma. 1])**.** *Consider an SCM over $\mathbf{V}$ induce observational distribution $P(\mathbf{V})$ and diagram $\mathcal{G}$. Suppose A_2 takes input as A_1. Then $\sum_{a_1} P(A_{1[\mathbf{b}_1]}, A_{2[\mathbf{b}_2]}, ...)$ is identifiable if and only if $P(A_{1[\mathbf{b}_1]}, A_{2[\mathbf{b}_2]}, ...)$ is identifiable.*

Now, we are ready to proceed to the proof of Thm. 1.

Theorem 1 (Graphical Criterion)**.** *Consider GCP models that employ a set of features $\mathbf{T}$ as a predictor of the label. $\Omega_{\text{GCP}(\mathbf{T})}$ is causally interpretable w.r.t. a query $Q(\mathbf{W})$ if and only if $\mathbf{T} \subseteq \mathbf{W} \cup ND(\mathbf{W})$.*

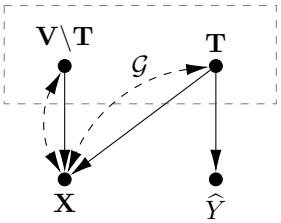


Figure 8: Diagrams used in the proof of Thm. 1.

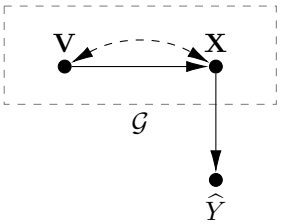


Figure 9: Diagrams used in the proof of Prop. 1.

Proof. According to Defs 2, 3 and 8, this is equivalent to prove that query $P(\widehat{y}_{\mathbf{w}'} \mid \mathbf{x})$ is identifiable iff $\mathbf{T} \subseteq ND(\mathbf{W}) \cup \mathbf{W}$ given the observational distribution $P(\mathbf{V}, \mathbf{X}, \widehat{Y})$ and the diagram $\mathcal{G}^{\text{Aug}}$ over $\{\mathbf{V}, \mathbf{X}, \widehat{Y}\}$ (shown in Fig. 8). To illustrate, the diagram $\mathcal{G}$ over $\mathbf{V}$ is an arbitrary given DAG; for any $V_i \in \mathbf{V}$, V_i point to X and bi-directed connected to X; only a subset $\mathbf{T} \subseteq \mathbf{V}$ point to $\widehat{Y}$. Denote $\mathbf{Z} = \mathbf{T} \backslash \mathbf{W}$.

$$
\begin{aligned}
& P(\widehat{y}_{\mathbf{w}'} \mid \mathbf{x}) \\
&= \sum_{\mathbf{v}} P(\widehat{y}_{\mathbf{w}'} \mid \mathbf{v}, \mathbf{x}) P(\mathbf{v} \mid \mathbf{x}) && \text{(summing over } \mathbf{V}\text{)} \\
&= \sum_{\mathbf{v}, \mathbf{t}''} P(\widehat{y}_{\mathbf{w}'} \mid \mathbf{t}''_{\mathbf{w}'}, \mathbf{v}, \mathbf{x}) P(\mathbf{t}''_{\mathbf{w}'} \mid \mathbf{v}, \mathbf{x}) P(\mathbf{v} \mid \mathbf{x}) && \text{(summing over } \mathbf{T}_{\mathbf{w}'} \text{ in } \mathcal{M}_{\mathbf{w}'} \text{world)} \\
&= \sum_{\mathbf{v}, \mathbf{t}''} P(\widehat{y}_{\mathbf{w}'} \mid \mathbf{t}''_{\mathbf{w}'}) P(\mathbf{t}''_{\mathbf{w}'} \mid \mathbf{v}, \mathbf{x}) P(\mathbf{v} \mid \mathbf{x}) && (\widehat{Y}_{\mathbf{w}'} \perp \{\mathbf{V}, \mathbf{X}\} \mid \mathbf{T}_{\mathbf{w}'}) \quad (6) \\
&= \sum_{\mathbf{v}, \mathbf{z}''} P(\widehat{y}_{\mathbf{w}'} \mid \mathbf{z}''_{\mathbf{w}'}) P(\mathbf{z}''_{\mathbf{w}'} \mid \mathbf{v}, \mathbf{x}) P(\mathbf{v} \mid \mathbf{x}) && \text{(consistency)} \quad (7) \\
&= \sum_{\mathbf{v}, \mathbf{z}''} P(\widehat{y} \mid \mathbf{z}'', \mathbf{w}') P(\mathbf{z}''_{\mathbf{w}'} \mid \mathbf{v}, \mathbf{x}) P(\mathbf{v} \mid \mathbf{x}) && \text{(do-calculus [25])} \quad (8)
\end{aligned}
$$

Eq. (6) holds since $\widehat{Y}_{\mathbf{w}}$ are independent with $\mathbf{X}$ and $\mathbf{V}$ since all parents of $\widehat{Y}_{\mathbf{w}'}$ (which is $\mathbf{T}_{\mathbf{w}'}$) are conditioned on. Eq. (7) holds since the $\mathbf{T} \cap \mathbf{W}$ should be consistent with the intervened value in $\mathbf{w}$ (and the remaining variables $\mathbf{Z}$ in $\mathbf{T}$ taking $\mathbf{z}''$. Eq. (8) holds due to $\widehat{Y} \perp \mathbf{W} \mid \mathbf{T}$ in $\mathcal{G}_{\underline{\mathbf{W}}}$, where $\mathcal{G}_{\underline{\mathbf{W}}}$ is the graph removing outgoing edge of $\mathbf{W}$. Using do-calculus, we have:

$$
P(\widehat{y}_{\mathbf{w}'} \mid \mathbf{z}''_{\mathbf{w}'}) = P(\widehat{y} \mid \mathbf{z}'', \mathbf{w}'). \tag{9}
$$

We will prove that Eq. (8) is identifiable if and only if $\mathbf{T} \subseteq ND(\mathbf{W}) \cup \mathbf{W}$, which is equivalent to prove Eq. (8) is identifiable iff $\mathbf{Z} \subseteq ND(\mathbf{W})$ since $\mathbf{Z} = \mathbf{T} \setminus \mathbf{W}$. According to Eq. (8), the only undermined term is $P(\mathbf{z}''_{\mathbf{w}'} \mid \mathbf{v}, \mathbf{x})$. Since $\mathbf{V}$ and $\mathbf{X}$ are bi-directly connected, Lemma 1 suggests $P(\mathbf{z}''_{\mathbf{w}'} \mid \mathbf{v}, \mathbf{x})$ is identifiable iff $\mathbf{Z} \subseteq ND(\mathbf{W})$. Then, $P(\widehat{y} \mid \mathbf{z}'', \mathbf{w}') P(\mathbf{z}''_{\mathbf{w}'} \mid \mathbf{v}, \mathbf{x}) P(\mathbf{v} \mid \mathbf{x})$ is identifiable iff $\mathbf{Z} \subseteq ND(\mathbf{W})$. According to Lemma 2, Eq. (8) is identifiable iff $\mathbf{T} \subseteq ND(\mathbf{W}) \cup \mathbf{W}$. □

Proposition 1 (Non-interpretability of BP)**.** *For any latent causal graph $\mathcal{G}_{\mathbf{V}}$, Ω_{BP} is not causally interpretable w.r.t. $Q(\mathbf{W})$ for any $\mathbf{W} \subseteq \mathbf{V}$.*

Proof. Since the observational $P(\mathbf{X})$ is identifiable, we will prove that $P(\widehat{y}_{\mathbf{w}'}, \mathbf{x})$ is not identifiable given a blackbox model structure and observational distribution $P(\mathbf{X}, \widehat{Y}, \mathbf{V})$.

$$
\begin{aligned}
& P(\widehat{y}_{\mathbf{w}'}, \mathbf{x}) \\
=& \sum_{\mathbf{x}'} P(\widehat{y}_{\mathbf{w}'}, \mathbf{x}'_{\mathbf{w}'}, \mathbf{x}) && \text{(summing over } \mathbf{X}_{\mathbf{w}'}) \\
=& \sum_{\mathbf{x}'} P(\widehat{y}_{\mathbf{x}'}) P(\mathbf{x}'_{\mathbf{w}'}, \mathbf{x}) && \text{(consistency and } \widehat{Y}_{\mathbf{x}'} \perp \{\mathbf{X}, \mathbf{X}_{\mathbf{w}'}\}) \\
=& \sum_{\mathbf{x}'} P(\widehat{y}_{\mathbf{x}'}) \sum_{\mathbf{w}} P(\mathbf{x}'_{\mathbf{w}'}, \mathbf{x}, \mathbf{w}) && \text{(summing over } \mathbf{W})
\end{aligned}
\tag{10}
$$

$P(\mathbf{x}'_{\mathbf{w}'}, \mathbf{x}, \mathbf{w})$ is not identifiable according to Lemma 1. Then Lemma 2 suggests that $P(\widehat{y}_{\mathbf{w}'}, \mathbf{x})$ is not identifiable. □

Theorem 2 (Uniqueness of Maximal T-Admissible Set)**.** *For the queries $Q(\mathbf{W}_\star)$, a maximal T-admissible set is unique and can be written as:*

$$\textit{Max-T-Ad}(\mathbf{W}_\star) = \cap_{\mathbf{W}_i \in \mathbf{W}_\star} \left(\mathbf{W}_i \cup ND(\mathbf{W}_i)\right). \tag{4}$$

Also, $\mathbf{T} \in$ T-Ad$(\mathbf{W}_\star)$ if and only if $\mathbf{T} \subseteq$ Max-T-Ad$(\mathbf{W}_\star)$.

Proof. (i) First, we will show that $\mathbf{S} := \cap_{\mathbf{W}_i \in \mathbf{W}_\star} \left(\mathbf{W}_i \cup ND(\mathbf{W}_i)\right)$ is a T-admissible set w.r.t $Q(\mathbf{W}_\star)$. For each $\mathbf{W}_i \in \mathbf{W}_\star$, we have

$$\cap_{\mathbf{W}_i \in \mathbf{W}_\star} \left(\mathbf{W}_i \cup ND(\mathbf{W}_i)\right) \subseteq \mathbf{W}_i \cup ND(\mathbf{W}_i).$$

Therefore, by Thm. 1, $\cap_{\mathbf{W}_i \in \mathbf{W}_\star} \left(\mathbf{W}_i \cup ND(\mathbf{W}_i)\right)$ is a T-admissible set w.r.t $Q(\mathbf{W}_i)$ for all $\mathbf{W}_i \in \mathbf{W}_\star$. Thus, we have $\mathbf{S} \in$ T-Ad$(\mathbf{W}_\star)$.

(ii) Now, we will show that $\mathbf{S}$ is a maximal T-admissible set w.r.t $\mathbf{W}_\star$. Suppose there exists $\mathbf{S}'$ such that $\mathbf{S}' \in$ T-Ad$(\mathbf{W}_\star)$ and $\mathbf{S}' \supsetneq \mathbf{S}$. Since $\mathbf{S}' \in$ T-Ad$(\mathbf{W}_\star)$, $\mathbf{S}' \in$ T-Ad$(\mathbf{W}_i)$ for all $\mathbf{W}_i \in \mathbf{W}_\star$. Hence,

$$\mathbf{S}' \subseteq \mathbf{W}_i \cup ND(\mathbf{W}_i) \quad \text{for all } \mathbf{W}_i \in \mathbf{W}_\star.$$

Therefore, $\mathbf{S}' \subseteq \cap_{\mathbf{W}_i \in \mathbf{W}_\star} \left(\mathbf{W}_i \cup ND(\mathbf{W}_i)\right) = \mathbf{S}$, which contradicts $\mathbf{S}' \supsetneq \mathbf{S}$. Therefore, $\mathbf{S}$ is a maximal T-admissible set w.r.t $\mathbf{W}_\star$.

(iii) Now, we will show that $\mathbf{S}$ is a unique maximal T-admissible set. Suppose there exists another maximal T-admissible set $\mathbf{S}'$. Since $\mathbf{S}' \in$ T-Ad$(\mathbf{W}_\star)$, we have $\mathbf{S}' \subseteq \mathbf{S}$ by the same reason in (ii). If $\mathbf{S}' \subsetneq \mathbf{S}$, then it contradicts that $\mathbf{S}'$ is a maximal T-admissible set, since $\mathbf{S}$ is a T-admissible set. Therefore, we have $\mathbf{S} = \mathbf{S}'$. In other words, a maximal T-admissible set is unique and can be written as Max-T-Ad$(\mathbf{W}_\star) = \cap_{\mathbf{W}_i \in \mathbf{W}_\star} \left(\mathbf{W}_i \cup ND(\mathbf{W}_i)\right)$.

(iv) Now, we will show that $\mathbf{T} \in$ T-Ad$(\mathbf{W}_\star)$ if and only if $\mathbf{T} \subseteq$ Max-T-Ad$(\mathbf{W}_\star)$. Suppose $\mathbf{T} \in$ T-Ad$(\mathbf{W}_\star)$. Then, by (ii), we have $\mathbf{T} \subseteq \cap_{\mathbf{W}_i \in \mathbf{W}_\star} \left(\mathbf{W}_i \cup ND(\mathbf{W}_i)\right)$. Also, we showed that Max-T-Ad$(\mathbf{W}_\star) = \cap_{\mathbf{W}_i \in \mathbf{W}_\star} \left(\mathbf{W}_i \cup ND(\mathbf{W}_i)\right)$. Therefore, we have $\mathbf{T} \subseteq$ Max-T-Ad$(\mathbf{W}_\star)$. Now, suppose that $\mathbf{T} \subseteq$ Max-T-Ad$(\mathbf{W}_\star)$. We have $\mathbf{T} \subseteq \cap_{\mathbf{W}_i \in \mathbf{W}_\star} \left(\mathbf{W}_i \cup ND(\mathbf{W}_i)\right)$, and thus, $\mathbf{T} \subseteq \mathbf{W}_i \cup ND(\mathbf{W}_i)$ for all $\mathbf{W}_i \in \mathbf{W}_\star$. Therefore, $\mathbf{T} \in$ T-Ad$(\mathbf{W}_i)$ for all $\mathbf{W}_i \in \mathbf{W}_\star$, and thus, $\mathbf{T} \in$ T-Ad$(\mathbf{W}_\star)$. □

Theorem 3 (Closed Form)**.** *If $\Omega_{\text{GCP}(\mathbf{T})}$ is causally interpretable w.r.t. $Q(\mathbf{W})$, the following holds:*

$$P(\widehat{Y}_{\mathbf{w}'} \mid \mathbf{x}) = \sum_{\mathbf{t}} P(\widehat{Y} \mid \mathbf{w}' \cap \mathbf{T}, \mathbf{t} \setminus \mathbf{W}) P(\mathbf{t} \mid \mathbf{x}). \tag{5}$$

Proof. From Eq. (8), we have

$$P(\widehat{y}_{\mathbf{w}} \mid \mathbf{x}) = \sum_{\mathbf{v}, \mathbf{z}''} P(\widehat{y} \mid \mathbf{z}'', \mathbf{w}') P(\mathbf{z}''_{\mathbf{w}'} \mid \mathbf{v}, \mathbf{x}) P(\mathbf{v} \mid \mathbf{x}). \tag{11}$$

Note that this equation is identifiable if only if $\mathbf{Z} \subseteq \mathbf{W} \cup ND(\mathbf{W})$. Then

$$
\begin{aligned}
&= \sum_{\mathbf{v},\mathbf{z}''} P(\widehat{y} \mid \mathbf{z}'', \mathbf{w}') P(\mathbf{z}''_{\mathbf{w}'} \mid \mathbf{v}, \mathbf{x}) P(\mathbf{v} \mid \mathbf{x}) \\
&= \sum_{\mathbf{v},\mathbf{z}''} P(\widehat{y} \mid \mathbf{z}'', \mathbf{w}') \mathbf{1}[\mathbf{z}'' = \mathbf{v}] P(\mathbf{v} \mid \mathbf{x}) && \text{(Lemma. 1)} \\
&= \sum_{\mathbf{v}} P(\widehat{y} \mid \mathbf{t} \setminus \mathbf{w}, \mathbf{w}') P(\mathbf{v} \mid \mathbf{x}) && \text{(where } \mathbf{z} = (\mathbf{t} \setminus \mathbf{w}) \in \mathbf{v}) \\
&= \sum_{\mathbf{v}} P(\widehat{y} \mid \mathbf{t} \setminus \mathbf{w}, \mathbf{w}' \cap \mathbf{t}) P(\mathbf{v} \mid \mathbf{x}) && (\mathbf{Y} \perp \mathbf{W} \setminus \mathbf{T} \mid \mathbf{T}) \\
&= \sum_{\mathbf{t}} P(\widehat{y} \mid \mathbf{t} \setminus \mathbf{w}, \mathbf{w}' \cap \mathbf{t}) P(\mathbf{t} \mid \mathbf{x}). && (12)
\end{aligned}
$$

This conclude $P(\widehat{Y}_{\mathbf{w}} \mid \mathbf{x}) = \sum_{\mathbf{t}} P(\widehat{Y} \mid \mathbf{w}' \cap \mathbf{T}, \mathbf{t} \setminus \mathbf{W}) P(\mathbf{t} \mid \mathbf{x})$ since Eq. 12 holds for any $\mathbf{t}, \mathbf{w}$. □

Theorem 4 (Causal Interpretability-Accuracy Trade-Off)**.** *The following holds:*
(i) If $\mathbf{T}_1 \subseteq \mathbf{T}_2$, then W-Ad$(\mathbf{T}_2) \subseteq$ W-Ad$(\mathbf{T}_1)$.
(ii) If $\mathbf{W}^1_\star \subseteq \mathbf{W}^2_\star$, then Max-T-Ad$(\mathbf{W}^2_\star) \subseteq$ Max-T-Ad$(\mathbf{W}^1_\star)$.

Proof. (i) Let $\mathbf{T}_1 \subseteq \mathbf{T}_2$. Suppose $\mathbf{W} \in$ W-Ad$(\mathbf{T}_2)$. By Def. 6 and Thm. 1, we have

$$\mathbf{T}_2 \subseteq \mathbf{W} \cup ND(\mathbf{W}).$$

Since $\mathbf{T}_1 \subseteq \mathbf{T}_2$, it follows that $\mathbf{T}_1 \subseteq \mathbf{W} \cup ND(\mathbf{W})$. Therefore, by Def. 6 and Thm. 1, $\mathbf{W} \in$ W-Ad$(\mathbf{T}_1)$. Thus, for all $\mathbf{W} \in$ W-Ad$(\mathbf{T}_2)$, we have $\mathbf{W} \in$ W-Ad$(\mathbf{T}_1)$. Hence, we have

$$\text{W-Ad}(\mathbf{T}_2) \subseteq \text{W-Ad}(\mathbf{T}_1).$$

(ii) Let $\mathbf{W}^1_\star \subseteq \mathbf{W}^2_\star$. Then, we have

$$\cap_{\mathbf{W}_i \in \mathbf{W}^2_\star} \left(\mathbf{W}_i \cup ND(\mathbf{W}_i)\right) \subseteq \cap_{\mathbf{W}_i \in \mathbf{W}^1_\star} \left(\mathbf{W}_i \cup ND(\mathbf{W}_i)\right).$$

Therefore, we have Max-T-Ad$(\mathbf{W}^2_\star) \subseteq$ Max-T-Ad$(\mathbf{W}^1_\star)$ by Thm. 2. □

A.3 Additional Examples

The following example illustrates how GCP and CP models compare.

Example 5 (GCP)**.** *Consider the generative process of observed concepts $\mathbf{V}_0 = \{F, S, C\}$ and the image $\mathbf{X}$, as in Ex. 1 (BP model) and Ex. 2 (CP model). Consider a GCP model $\mathcal{M}_{GCP} = \langle \mathbf{U} = \{U_F, U_S, U_{C_1}, U_{C_2}, \mathbf{U_X}\}, \{\{F, S, C\}, \mathbf{X}, \widehat{Y}\}, \mathcal{F}^{GCP}, P^{GCP}(\mathbf{U})\rangle$, where*

$$
\mathcal{F}^{GCP} = \begin{cases}
F \leftarrow U_F \oplus U_S \\
S \leftarrow U_S \\
C \leftarrow (\neg S \wedge U_{C_1}) \oplus (S \wedge U_{C_2}) \\
\mathbf{X} \leftarrow f_{\mathbf{X}}(F, S, C, \mathbf{U_X}) \\
\widehat{Y} \leftarrow f^{GCP}_{\widehat{Y}}(S, F)
\end{cases} \tag{13}
$$

and $P^{GCP}(\mathbf{U})$ is equal to $P^{CP}(\mathbf{U})$ in Ex. 2. The causal diagram induced by GCP model $\mathcal{M}_{GCP}$ is shown in Fig. 2c. To illustrate, instead of predicting the label based on pixels in images $\mathbf{X}$ (BP models) or all observed features $\{F, S, C\}$ (CP models), GCP model makes a prediction using a selected subset of features $\mathbf{T} = \{S, F\}$ (i.e., smiling and gender) in this case. ■

The following example illustrates the case where the GCP model is causal interpretable.

Example 6 (Continued from Ex. 5)**.** *Consider Ω_{CP} in Ex. 4. Thm. 1 suggests Ω_{CP} is not interpretable w.r.t. to query $Q(S)$ $P(Y_{S=0} \mid \mathbf{X})$. This is because $C \in De(S)$, where*

F	S	C	$P(F, S, C) = 1$
0	0	0	0.168
0	0	1	0.072
0	1	0	0.096
0	1	1	0.144
1	0	0	0.112
1	0	1	0.048
1	1	0	0.144
1	1	1	0.216

Table 1: Probability table in Ex. 6.

$\mathbf{W} = \{S\}$, *i.e., the prediction of* $\widehat{Y}$ *is made based on* C*, a descendant of* S*. In contrast,* $\Omega_{\text{GCP}(\{S,F\})}$ *in Ex. 5 is said to be causally interpretable w.r.t. to query* $P(Y_{S=0} \mid \mathbf{X})$ *since* $f^{GCP}_{\widehat{Y}}$ *only takes* $\mathbf{T} = \{S, F\} \subseteq S \cup ND(S)$ *as input. To illustrate, let us consider the GCP model* $\mathcal{M}_{GCP}$ *in Ex. 5. Similar to Examples 3 and 4, let the observational quantity* $P(F = 0, S = 1, C = 1 \mid \mathbf{X} = \mathbf{x}) = 1$ *and let* $f_{\widehat{Y}}$ *be:*

$$\widehat{Y} \leftarrow f^{GCP}_{\widehat{Y}}(S, F) = \mathbf{1}[S + F > 0.5]. \tag{14}$$

Now, consider another GCP model

$$\mathcal{M}'_{GCP} = \langle \mathbf{U}' = \{U'_F, U'_{S_1}, U'_{S_1}, U'_{C_1}, U'_{C_2}, \mathbf{U}'_{\mathbf{X}}\}, \{\{F, S, C\}, \mathbf{X}, \widehat{Y}\}, \mathcal{F}^{GCP'}, P^{GCP'}(\mathbf{U})\rangle, \tag{15}$$

where

$$\mathcal{F}^{GCP'} = \begin{cases} F \leftarrow U'_F \\ S \leftarrow ((\neg U'_F) \wedge U'_{S_1}) \oplus (U'_F \wedge U'_{S_2}) \\ C \leftarrow (\neg S \wedge U'_{C_1}) \oplus (S \wedge U'_{C_2}) \\ \mathbf{X} \leftarrow f_{\mathbf{X}}(F, S, C, \mathbf{U}_{\mathbf{X}}) \\ \widehat{Y} \leftarrow \mathbf{1}[S + F > 0.5] \end{cases} \tag{16}$$

and $P(U'_F = 1) = 0.52, P(U'_{S_1} = 1) = 0.5, P(U'_{S_2} = 1) = 9/13, P(U'_{C_1} = 1) = 0.5, P(U'_{C_2} = 1) = 0.6$*. It is verifiable that* $P^{\mathcal{M}_{GCP}}(\mathbf{V}) = P^{\mathcal{M}'_{GCP}}(\mathbf{V})$ *as shown in Table 1. Since* $f_{\widehat{Y}}$ *is the same in both* $\mathcal{M}_{GCP}$ *and* $\mathcal{M}'_{GCP}$*,* $P^{\mathcal{M}_{GCP}}(\mathbf{V}, \widehat{Y}) = P^{\mathcal{M}'_{GCP}}(\mathbf{V}, \widehat{Y})$*. Let the distribution of* $\mathbf{U}_{\mathbf{X}}$ *satisifies that* $P^{\mathcal{M}_{GCP}}(\mathbf{V}, \mathbf{X}, \widehat{Y}) = P^{\mathcal{M}'_{GCP}}(\mathbf{V}, \mathbf{X}, \widehat{Y})$*.* $\mathcal{M}'_{CP}$ *is compatible the graphical constraints induced by the model in Fig. 2b. Notice that* f'_F, f'_S, f'_C *in* $\mathcal{M}'_{GCP}$ *are totally different to* f_F, f_S, f_C *in* $\mathcal{M}_{GCP}$*. For the first GCP model* $\mathcal{M}_{GCP}$*,*

$$P^{\mathcal{M}_{GCP}}(\widehat{Y}_{S=0} = 1 \mid \mathbf{X} = \mathbf{x}) = P^{\mathcal{M}_{GCP}}(F_{S=0} = 1 \mid F = 0, S = 1, C = 1) = 0.$$

Similarly, for the second GCP model $\mathcal{M}'_{GCP}$*,*

$$P^{\mathcal{M}'_{GCP}}(\widehat{Y}_{S=0} = 1 \mid \mathbf{X} = \mathbf{x}) = P^{\mathcal{M}'_{GCP}}(C_{S=0} = 1 \mid F = 0, S = 1, C = 1) = 0.$$

This shows that the two GCP models are consistent with the query. In other words, if one uses the features $\{S, F\}$ *to predict* $\widehat{Y}$ *, the model architecture in Fig. 2c is guaranteed to provide a unique answer for the counterfactual question "What would the attractiveness prediction be had the person not smiled?" (i.e.,* $P(Y_{S=0} \mid \mathbf{X})$*). Then one can trust the counterfactual quantities induced by any model with this architecture.* ■

B Experiments

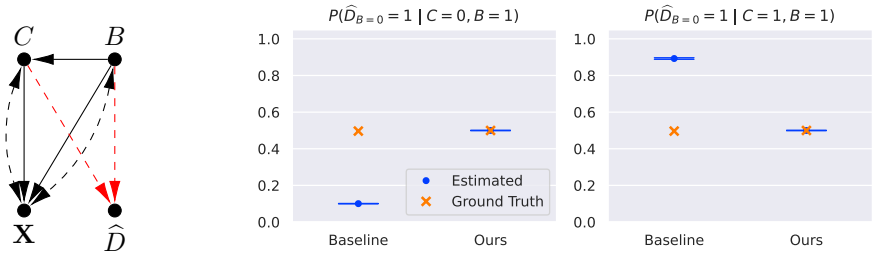
In this section, we describe the details for the experiments and provide additional experimental results.

B.1 Dataset

B.1.1 BarMNIST

For BarMNIST experiment discussed in Sec. 4.1, the data generating process is as follows:

$$\mathcal{F} = \begin{cases} D \leftarrow U_D \\ C \leftarrow U_D \oplus U_C \\ B \leftarrow (U_{B_1} \wedge D) \oplus (U_{B_1} \wedge U_{B_2}) \oplus ((\neg U_{B_1}) \wedge U_{B_2}) \\ \mathbf{X} \leftarrow f_{\mathbf{X}}(B, D, C, \mathbf{U}_{\mathbf{X}}) \\ Y \leftarrow \big((D \oplus C) \vee B\big) \oplus U_Y, \end{cases} \tag{17}$$



(a) Causal diagram.

(b) Estimation of counterfactuals.

Figure 10: (a) Causal diagram of GCP models. Red arrows represent the possible usage for predicting the label. (b) Estimation of counterfactual queries. Blue dots and orange marks denote estimation of counterfactual queries and ground truth value, respectively.

the exogenous variables $U_D, U_C, U_{B_1}, U_{B_2}, U_{B_3}, U_Y$ are independent binary variables, and $P(U_D = 1) = 0.5, P(U_C = 1) = 0.4, P(U_{B_1} = 1) = 0.9, P(U_{B_2} = 1) = 1/18, P(U_{B_3} = 1) = 0.5, P(U_Y = 1) = 0.1$.

Following this process, we generated 60,000 images and corresponding labels, where each image is annotated with 3 binary features, i.e., bar (B), color (C), and digit (D). Here, $D = 0$ represents the digits from 0 to 4 and $D = 1$ represents the digits from 5 to 9.

B.1.2 CelebA

CelebA dataset [19] contains 202,599 celebrity facial images, where each image is annotated with 40 different attributes. In our experiments, we used the attribute "attractiveness" as the label, where the label and all other features are binary.

B.2 Experimental Details

In BarMNIST, we used ResNet18 for the feature extractor. For the classifier, we used a three-layer MLP with the hidden dimension of 32 and leakyrelu activation. We set the batch size to 1024 and trained the models for 100 epoch. We used Adam optimizer with a learning rate of 0.0003.

In CelebA, we used ResNet34 for the feature extractor and used linear classifier. We set the batch size to 512 and trained the models for 100 epochs. We used SGD optimizer with the learning rate of 0.001. We resized the image with center crop into 64×64 for training.

For the training of our model and baselines, we used binary classification loss for both the feature extractor and the classifier, where they are trained simultaneously in an end-to-end manner. All experimental results are averaged over 5 independent runs. We report a standard error as the error bar in Figs. 6, 10 and 11. All experiments are conducted on a single NVIDIA A100 GPU. For the implementation, we utilized publicly available code from Espinosa Zarlenga et al. [6]. We used GPT-4o to generate the counterfactual images shown in Figs. 7 and 11 to provide an intuitive understanding of the counterfactual questions.

B.3 Additional Experimental Results

B.3.1 BarMNIST

To validate our theory with a different graph structure, we consider a causal diagram in Fig. 10a where the goal is to predict the digit D from the image. The data generating process is as follows:

$$\mathcal{F} = \begin{cases} B \leftarrow U_B \\ C \leftarrow B \vee U_{C_1} \oplus U_{C_2} \\ D \leftarrow (B \vee C) \oplus U_D \\ \mathbf{X} \leftarrow f_{\mathbf{X}}(B, D, C, \mathbf{U}_{\mathbf{X}}), \end{cases} \tag{18}$$

where the exogenous variables $U_B, U_{C_1}, U_{C_2}, U_D$ are independent binary variables, where $P(U_B = 1) = 0.6, P(U_{C_1} = 1) = 0.5, P(U_{C_2} = 1) = 0.1, P(U_D = 1) = 0.1$.

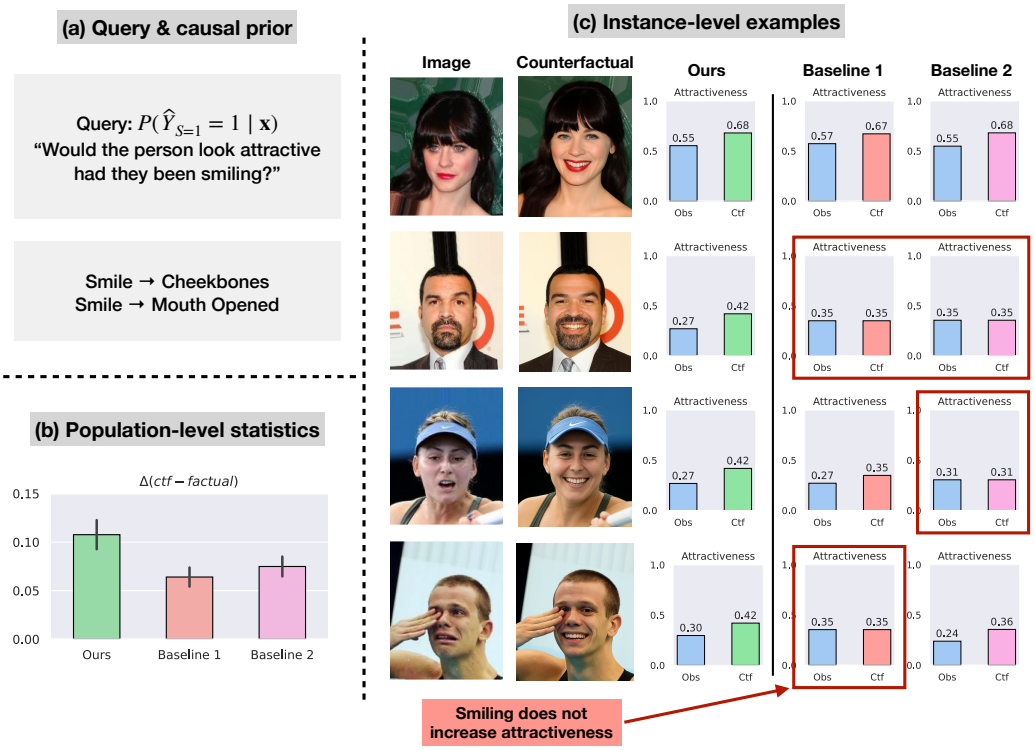


Figure 11: (a) We examine the prediction of the models under counterfactual condition. We use causal prior knowledge that smiling has causal effects on the features "cheekbones" and "opened mouth". (b) Average difference between the estimated counterfactual prediction and the prediction on the observed (factual) image. (c) Qualitative examples for our model and baselines.

The baseline model uses the features B and C for predicting the label, and our model uses B for making a prediction. Our theory (Thm. 1) suggests that our model is causally interpretable, but not the baseline which uses C, a descendant of B. We compare our model and baselines for estimating the counterfactual prediction of the model, where the query is to change the bar, i.e., $P(\widehat{D}_{B=0} \mid \mathbf{x})$.

Fig. 10b illustrates the estimation of counterfactual queries (blue dots) and ground truth values (orange marks). This shows that our model correctly estimates counterfactual queries. In contrast, the estimation of the baseline significantly differs from the ground truth. This corroborates our theory that our estimation can properly interpret the counterfactual behavior of the causally interpretable models, but it is not possible for non-interpretable ones.

B.3.2 CelebA

Here, we provide a detailed analysis of CelebA experiments in Sec. 4.2. Fig. 11-(a) illustrates the counterfactual question and causal prior we utilized to construct our model. Specifically, we leverage the common-sense knowledge that smiling has direct causal influence to the features "cheekbones" and "opened mouth". To construct our model, we choose features that are non-descendants of smiling, specifically "smile" and "gender" as feature set $\mathbf{V}$. Baselines include descendant features. In Fig. 11, baseline 1 uses the features "smiling", "gender", and "cheekbones" and baseline 2 uses the features "smiling", "gender", "cheekbones", and "opened mouth".

Fig. 11-(b) shows the average difference between the estimated counterfactual prediction and the prediction on the observed image. Fig. 11-(c) shows qualitative examples comparing our method and baselines. The first column in Fig. 11-(c) shows the input image, and the second column illustrates the counterfactual image, as a reference to provide a better understanding of the counterfactual query.

The theory suggests that a causally interpretable model can properly estimate its prediction under counterfactual conditions. As shown in Fig. 11-(b) and (c), our model, which is causally interpretable, consistently increases the attractiveness across the instances, which is also aligned with human reasoning. In contrast, as illustrated in Fig. 11-(c), the estimation of the baselines (which use similar feature set as ours) shows that smiling often does not increase attractiveness (red boxes). In fact, our theory suggests that it is not possible to interpret the counterfactual behavior of non-interpretable models using only observational data, and any attempts to estimate it would lead to inconsistent results.

C Additional Discussions, Limitations, and Future Work

Estimation of the concepts. In the closed-form formula in Eq. (5), the concepts $\mathbf{W}$ and $\mathbf{T}$ are ground-truth concepts. Since the labels of ground-truth concepts are available, one can estimate $P(\mathbf{T} \mid \mathbf{X})$ over ground truth concept $\mathbf{T}$. For clarity, let us denote this estimated distribution as $\widehat{P}(\mathbf{T} \mid \mathbf{X})$. In the prediction stage, the true concepts $\mathbf{W}$ and $\mathbf{T}$ of an image instance $\mathbf{X}$ are not given directly. Instead, the predicted concepts $\mathbf{W}$ and $\mathbf{T}$ are sampled through the estimated $\widehat{P}(\mathbf{T} \mid \mathbf{X})$. When $\widehat{P}(\mathbf{T} \mid \mathbf{X})$ is accurate, the sampled (predicted) concepts are expected to align closely with the ground-truth concepts. However, if the estimation has an error, the predicted concepts may deviate from the true ones, and this error will naturally propagate into the counterfactual evaluation via Eq. (5).

Our goal with this formulation is to formally characterize how these counterfactual quantities can be computed from the observational distribution under ideal conditions (i.e., accurate estimation). The challenge of robustly estimating $P(\mathbf{T} \mid \mathbf{X})$ from finite data is indeed fundamental and highly relevant to practice, but falls outside the scope of this work. Nevertheless, it would be a valuable direction for future investigation, particularly in light of ongoing research in counterfactual estimation within the causal inference literature [12, 13] and the importance of creating more interpretable methods in practice.

Causal graph. Our work reveals that understanding and harnessing causal relationships among the generative features are crucial for building interpretable models that can properly evaluate counterfactual questions. It is important to note that our framework only requires the causal prior on the descendants, and this is a much relaxed assumption compared to the conventional assumption in causal inference, where the full specification of the causal graph is needed [9, 18].

Real-world datasets. In real-world datasets, it is infeasible to evaluate the actual value of the counterfactual query because the underlying ground-truth data-generating process for real-world datasets is not given, specifically, the mechanisms of $\mathbf{V}$ are not known. For example, it is unknown how nature decides the generation process of human facial features. Due to this inevitable restriction, we thoroughly validated our theory in BarMNIST datasets (where we have the ground-truth SCM), including causal interpretability-accuracy tradeoff.

Still, our theory allows us to understand the interplay between causal interpretability and accuracy in real-world datasets. For example, given T-admissible set smiling, gender and the query "Would the person be attractive had they smiled?", if one wants to incorporate additional query "Would the person be attractive had they be a men?", we know the model using this $\mathbf{T}$ maintains causal interpretability w.r.t. both queries, and thus, would not compromise accuracy.